Title
Expert-augmented machine learning.

Journal
Proceedings of the National Academy of Sciences of the United States of America, 117(9)

ISSN
0027-8424

Authors
Gennatas, Efstathios D
Friedman, Jerome H
Ungar, Lyle H
et al.

Publication Date
2020-03-01

DOI
10.1073/pnas.1906831117

Peer reviewed

Expert-augmented machine learning

Efstathios D. Gennatas[a,1,2], Jerome H. Friedman[b], Lyle H. Ungar[c], Romain Pirracchio[d], Eric Eaton[c], Lara G. Reichmann[e], Yannet Interian[e], José Marcio Luna[f], Charles B. Simone II[g], Andrew Auerbach[h], Elier Delgado[i], Mark J. van der Laan[j], Timothy D. Solberg[a], and Gilmer Valdes[a]

[a]Department of Radiation Oncology, University of California, San Francisco, CA 94143; [b]Department of Statistics, Stanford University, Stanford, CA 94305; [c]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104; [d]Department of Anesthesia and Perioperative Care, University of California, San Francisco, CA 94143; [e]Data Institute, University of San Francisco, CA 94105; [f]Department of Radiation Oncology, University of Pennsylvania, Philadelphia, PA 19104; [g]Department of Radiation Oncology, New York Proton Center, New York, NY 10035; [h]Division of Hospital Medicine, University of California, San Francisco, CA 94143; [i]Innova Montreal, Inc., Montreal, QC J4W 2P2, Canada; and [j]Division of Biostatistics, University of California, Berkeley, CA 94720

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved January 22, 2020 (received for review April 29, 2019)

Machine learning is proving invaluable across disciplines. However, its success is often limited by the quality and quantity of available data, while its adoption is limited by the level of trust afforded by given models. Human vs. machine performance is commonly compared empirically to decide whether a certain task should be performed by a computer or an expert. In reality, the optimal learning strategy may involve combining the complementary strengths of humans and machines. Here, we present expert-augmented machine learning (EAML), an automated method that guides the extraction of expert knowledge and its integration into machine-learned models. We used a large dataset of intensive-care patient data to derive 126 decision rules that predict hospital mortality. Using an online platform, we asked 15 clinicians to assess the relative risk of the subpopulation defined by each rule compared to the total sample. We compared the clinician-assessed risk to the empirical risk and found that, while clinicians agreed with the data in most cases, there were notable exceptions where they overestimated or underestimated the true risk. Studying the rules with greatest disagreement, we identified problems with the training data, including one miscoded variable and one hidden confounder. Filtering the rules based on the extent of disagreement between clinician-assessed risk and empirical risk, we improved performance on out-of-sample data and were able to train with less data. EAML provides a platform for automated creation of problem-specific priors, which help build robust and dependable machine-learning models in critical applications.

machine learning | medicine | computational medicine

Machine-learning (ML) algorithms are proving increasingly successful in a wide range of applications but are often data inefficient and may fail to generalize to new cases. In contrast, humans are able to learn with significantly less data by using prior knowledge. Creating a general methodology to extract and capitalize on human prior knowledge is fundamental for the future of ML. Expert systems, introduced in the 1960s and popularized in the 1980s and early 1990s, were an attempt to emulate human decision-making in order to address artificial intelligence problems (1). They involved hard-coding multiple if–then rules laboriously designed by domain experts. This approach proved problematic because a very large number of rules was usually required, and no procedure existed to generate them automatically. In practice, such methods commonly resulted in an incomplete set of rules and poor performance. The approach fell out of favor and attention has since been focused mainly on ML algorithms requiring little to no human intervention. More recently, the Prognosis Research Strategy Partnership of the United Kingdom's Medical Research Council has published a series of recommendations to establish a framework for clinical predictive model development, which emphasize the important of human expert supervision of model training, validation, and updating (2, 3).

Learning algorithms map a set of features to an outcome of interest by taking advantage of the correlation structure of the data. The success of this mapping will depend on several factors, other than the amount of actual information present in the covariates (also known as features, also known as independent variables), including the amount of noise in the data, the presence of hidden confounders, and the number of available training examples. Lacking any general knowledge of the world, it is no surprise that current ML algorithms will often make mistakes that would appear trivial to a human. For example, in a classic study, an algorithm trained to estimate the probability of death from pneumonia labeled asthmatic patients as having a lower risk of death than nonasthmatics (4). While misleading, the prediction was based on a real correlation in the data: These patients were reliably treated faster and more aggressively, as they should, resulting in consistently better outcomes. Out of context, misapplication of such models could lead to catastrophic results (if, for example, an asthmatic patient was discharged prematurely or undertreated). In a random dataset collected to illustrate the widespread existence of confounders in medicine, it was found that colon cancer screening and abnormal breast findings were highly correlated to the risk of having a stroke, with no apparent clinical justification (5).

Significance

Machine learning is increasingly used across fields to derive insights from data, which further our understanding of the world and help us anticipate the future. The performance of predictive modeling is dependent on the amount and quality of available data. In practice, we rely on human experts to perform certain tasks and on machine learning for others. However, the optimal learning strategy may involve combining the complementary strengths of humans and machines. We present expert-augmented machine learning, an automated way to automatically extract problem-specific human expert knowledge and integrate it with machine learning to build robust, dependable, and data-efficient predictive models.

Author contributions: E.D.G., J.H.F., and G.V. designed research; E.D.G. and G.V. performed research; E.D.G., R.P., and L.G.R. contributed new reagents/analytic tools; E.D.G. analyzed data; and E.D.G., J.H.F., L.H.U., R.P., E.E., L.G.R., Y.I., J.M.L., C.B.S., A.A., E.D., M.J.v.d.L., T.D.S., and G.V. wrote the paper.

Competing interest statement: The editor, P.J.B., and one of the authors, M.J.v.d.L., are at the same institution (University of California, Berkeley).

This article is a PNAS Direct Submission.

Data deposition: Data collected from clinician experts along with code used for analysis has been deposited online at GitHub, https://github.com/egenn/EAML_MIMIC_ICUmortality.

[1]To whom correspondence may be addressed. Email: gennatas@stanford.edu.

[2]Present address: Department of Radiation Oncology, Stanford University, Stanford, CA 94305.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1906831117/-/DCSupplemental.

First published February 18, 2020.

Unfortunately, superior performance on a task as measured on test sets derived from the same empirical distribution, is often considered as evidence that real knowledge has been captured by a model. In a recent study, CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, investigators observed that a convolutional neural network (CNN) outperformed radiologists in overall accuracy (6). A subsequent study revealed that the CNN was basing some of its predictions on image artifacts that identified hospitals with higher prevalence of pneumonia or discriminated regular from portable radiographs (the latter is undertaken on sicker patients), while pathology present in the image was sometimes disregarded (7). It was also shown that performance declined when a model trained with data from one hospital was used to predict data from another (8).

Among the biggest challenges for ML in high-stakes applications like medicine is to automatically extract and incorporate prior knowledge that allows ML algorithms to generalize to new cases and learn with less data. In this study, we hypothesized that combining the extensive prior knowledge of causal and correlational physiological relationships that human experts possess with a machine-learned model would increase model generalizability, i.e., out-of-sample performance. We introduce expert-augmented machine learning (EAML), a methodology to automatically acquire problem-specific priors and incorporate them into an ML model. The procedure allows training models with 1) less data that are 2) more robust to changes in the underlying variable distributions and 3) resistant to performance decay with time. Rather than depending on hard-coded and incomplete rule sets, like the early expert systems did, or relying on potentially spurious correlations like current ML algorithms often do, EAML guides the acquisition of prior knowledge to improve the final ML model. We demonstrate the value of EAML using the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) dataset collected at the Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2012 and released by the PhysioNet team to predict mortality among intensive-care unit (ICU) patients (9, 10).

EAML Generates Problem-Specific Priors from Human Domain Experts

To automate the generation of problem-specific priors, we developed a multistep approach (see *Methods* and summary in Fig. 1). First, we trained RuleFit on the MIMIC-II ICU dataset collected at the BIDMC between 2001 and 2008 to predict hospital mortality using 17 demographic and physiologic input variables that are included in popular ICU scoring systems (11–13). This yielded 126 rules with nonzero coefficients. Using a 70%/30% training/test split on the 24,508 cases, RuleFit achieved a test set balanced accuracy of 74.4 compared to 67.3 for a Random Forest. Previously,

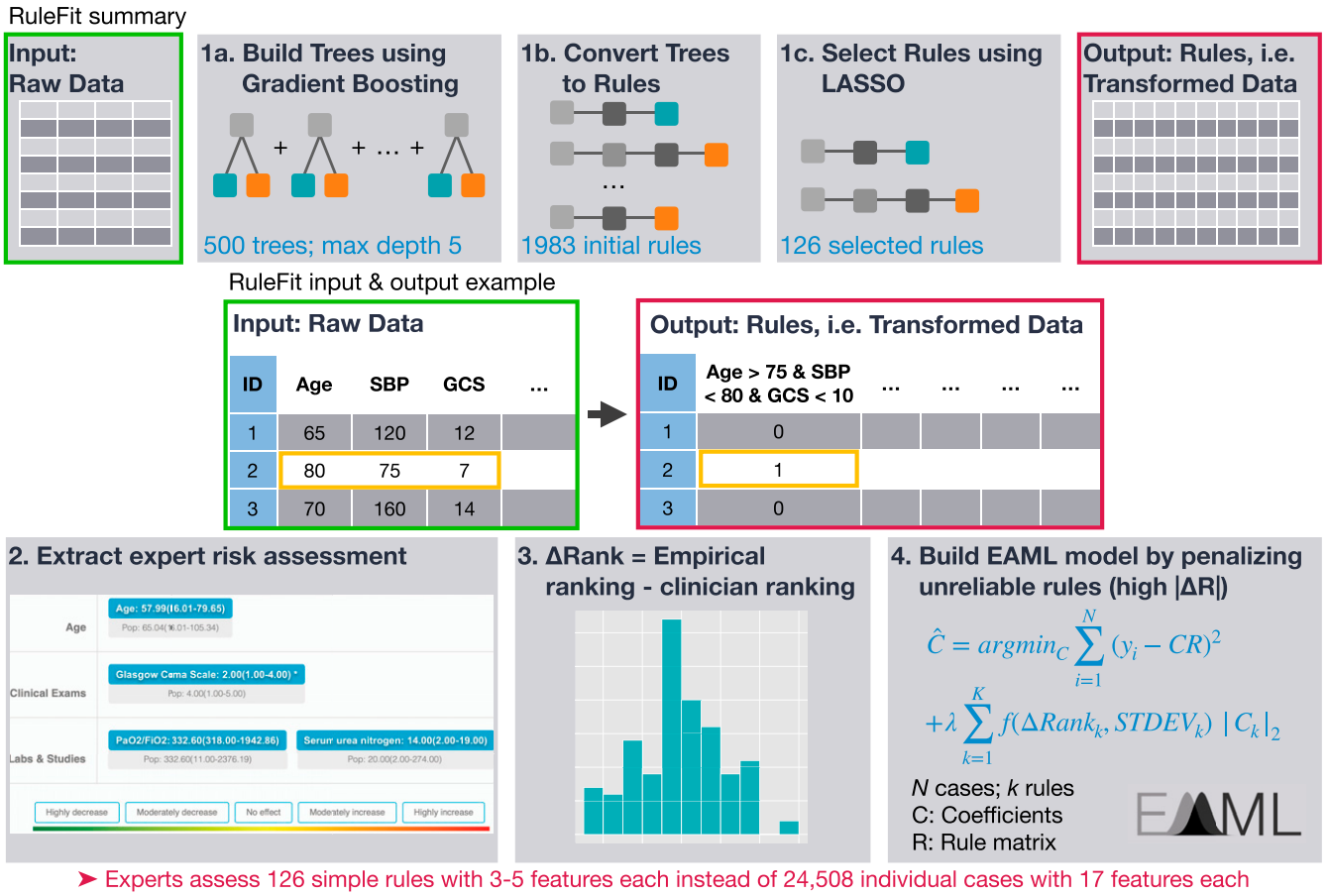

Fig. 1. Overview of the methods. RuleFit involves 1) training a gradient boosting model on the input data, 2) converting boosted trees to rules by concatenating conditions from the root node to each leaf node, and 3) training an L1-regularized (LASSO) logistic regression model. Each rule defines a subpopulation that satisfies all conditions in the rule. Clinician experts assess the mortality risk of the subpopulation defined by each rule compared to the whole sample on a web application. For each rule, delta ranking is calculated as the difference between the subpopulation's empirical risk as suggested by the data and the clinicians' estimate. A final model is trained by reducing the influence of those rules with highest delta ranking. This forms an efficient procedure where experts are asked to assess 126 simple rules of 3 to 5 variables each instead of assessing 24,508 cases with 17 variables each.

Random Forest had been found to be the top performer among a library of algorithms on the MIMIC-II dataset (14). Subsequently, a committee of 15 clinicians at the University of California, San Francisco, were asked to categorize the risk of the subpopulations defined by each rule compared to the general population without being shown the empirical risk (Fig. 2). On average, clinicians took 41 ± 19 min to answer 126 questions.

We calculated the average clinician assessment for each rule and ranked rules by increasing perceived risk, Rank_p. To check that we were successful in acquiring valid clinical information, we then binned the rules into five groups according to their ranking and plotted the empirical risk by group (Fig. 3). There is a monotonic relationship between the average clinicians' ranking of a rule and its empirical risk (mortality ratio), as expected.

Delta Rank Helps Discover Hidden Confounders

The mortality ratio of patients within the subpopulation defined by each rule was used to calculate the empirical risk ranking of the rules, Rank_e. The delta ranking was defined as $\Delta\text{Rank} = \text{Rank}_e - \text{Rank}_c$ and is a measure of clinicians' disagreement with the empirical data. The distribution of ΔRank is shown in *SI Appendix*, Fig. S1. We hypothesized that those rules where ΔRank was outside the 90% confidence interval were likely to indicate either that clinicians misjudged the risk of the given subpopulations or that hidden confounders were modifying the risk. This hypothesis is based on the fact that the rules were created by the ML model based on empirical risk, while clinicians were estimating risk of each subpopulation based on medical knowledge and experience. We first analyzed those rules where the empirical ranking was significantly lower than the clinicians' perceived ranking (Table 1, top). For rules 1, 3, and 4, clinicians estimated that patients with a lower heart rate (HR) and Glasgow Coma Scale (GCS) score below 13 (in the original scale) are at higher risk than that supported by the data. For rules 2 and 5, clinicians appeared to overestimate the mortality risk of old age. Although it is true that older patients are generally at higher risk (11, 12, 15), the data suggest that being over 80 y old does not automatically increase one's risk of death in the ICU, if their physiology is not otherwise particularly compromised. Finally, the last rule in Table 1, top, indicates the discovery of a hidden confounder: intubation status. Intubated patients, whose responsiveness could not be assessed, were assigned the lowest possible GCS score in the MIMIC dataset. This was confirmed in correspondence with the PhysioNet team. Such a score would normally suggest a gravely ill patient who is unresponsive to external stimuli. Because the intubation status had not been initially collected, we reconstructed the same group of patients using MIMIC-III data and verified the miscoding (10). Patients with a GCS less than 8 who are not intubated (n = 1,236) have a mortality risk of 0.28. Conversely, intubated patients (n = 6,493) have a much lower mortality ratio of 0.19. The fact that intubated patients have been assigned the lowest possible GCS in the MIMIC-II dataset has largely been ignored in the literature. It was briefly mentioned by the PhysioNet team in the calculation of the sequential organ failure assessment (previously known as sepsis-related organ failure assessment; SOFA) score in the MIMIC-III dataset (16).

Table 1, bottom, shows the top 5% of the rules where the experts' ranking is lower than the empirical ranking. Here we find that clinicians have underestimated the influence of high blood urea nitrogen (BUN) or high bilirubin (rules 7, 8, 10, and 11), although it is known that these variables affect mortality (17–19). The disagreement with the rules 9 and 12 allowed us to identify another important issue with the data: Clinicians assigned a lower risk to patients with high ratio of arterial oxygen partial pressure to fractional inspired oxygen (PaO_2/FiO_2) than is supported by empirical data. In MIMIC-II, 54% of patients had missing values for PaO_2/FiO_2. After imputation with the mean, they were assigned a value of 332.60, which is very close to the value used by the rules in Table 1, bottom (342.31 and 336.67, respectively). We discovered that PaO_2/FiO_2 values were not missing at random. In total, 94.2% of patients (n = 14,430) with missing values for PaO_2/FiO_2 were not intubated, while 60.35% of patients with values for PaO_2/FiO_2 were intubated. Patients that were not intubated and had a PaO_2/FiO_2 greater than 336.67 had a mortality ratio of 0.046, which would agree with the clinicians' assessment. In contrast, patients that were intubated and had a PaO_2/FiO_2 greater than 336.67 had a mortality ratio of 0.13. Since this is ∼60% of patients, they dominated the mortality risk on these rules (e.g., 0.10 for the last rule on Table 1, bottom). As such, clinicians are again estimating risk based on their understanding of the effects of PaO_2/FiO_2 on mortality, while the algorithm has learned the effect of a hidden confounder; intubated vs. not intubated. To confirm this, we predicted intubation status in MIMIC-III patients from the other covariates and achieved 97%

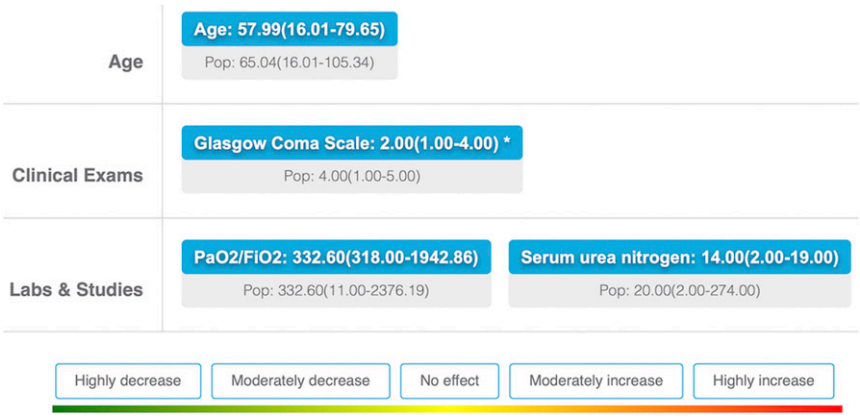

Fig. 2. Example of a rule presented to clinicians. Age, GCS (1, <6; 2, 6 to 8; 3, 9 to 10; 4, 11 to 13; 5, 14 to 15), ratio of oxygen blood concentration to fractional inspired oxygen concentration (PaO_2/FiO_2), and BUN concentration are the variables selected for this rule. The decision tree rules derived from gradient boosting, e.g., age ≤ 73.65 and GCS ≤ 4, were converted to the form "median (range)", e.g., age, 56.17 (16.01 to 73.65), for continuous variables and to the form "mode (included levels)" for categorical variables. Rules were presented in a randomized order, one at a time. The top line (blue box) displays the values for the subpopulation defined by the given rule. The bottom line (gray box) displays the values of the whole population. Participants were asked to assess the risk of belonging to the defined subpopulation compared to the whole sample using a five-point system: highly decrease, moderately decrease, no effect, moderately increase, and highly increase.

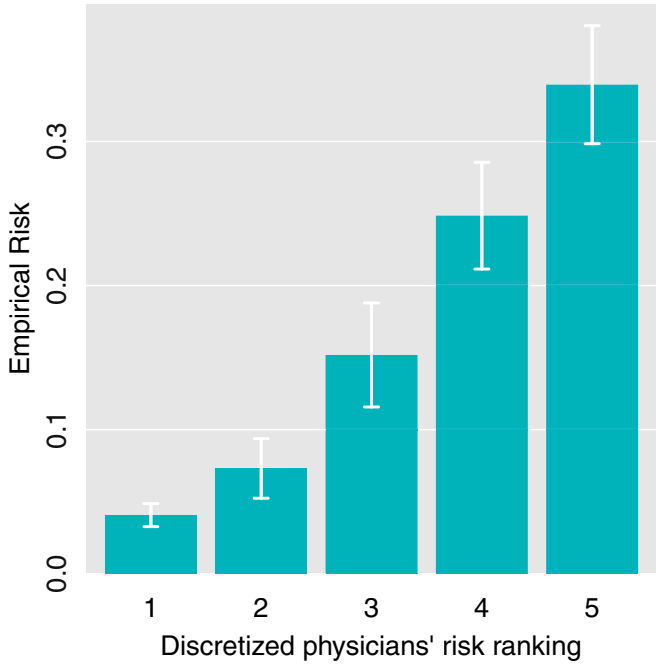

Fig. 3. Mortality ratio by average clinicians' risk ranking. Rules were binned into quintiles based on average clinicians' assessment. The mean empirical risk for each quintile was plotted. Error bars indicate 1.96 * SE.

mean accuracy using 10-fold cross-validation. This is especially troublesome because PaO_2/FiO_2 was selected by Random Forest as the most important variable in predicting mortality and was also selected as the most important variable driving clinicians' answers (Fig. 4). The underlying reason in each case is, however, very different, as the algorithm is using PaO_2/FiO_2 as a proxy of intubation while clinicians are answering based on their understanding of physiology.

EAML Improves Out-of-Sample Performance

The MIMIC dataset was well suited for us to test whether EAML can make models more robust to variable shifts or decay of accuracy with time. We built models combining clinicians' answers and the MIMIC-II dataset (collected from 2001 to 2008). We then evaluated these models on two sets of the MIMIC-III data: MIMIC-III1, which utilizes the same patients as in MIMIC-II but has different values of the input variables due to recoding of the underlying tables by the PhysioNet project, and MIMIC-III2 (collected from 2008 to 2012), which consists of new patients treated in the 4 y that followed the acquisition of MIMIC-II. Fig. 5*A* illustrates an example of a variable distribution change from MIMIC-II to MIMIC-III1 (i.e., on the same cases).

Fig. 5*B* illustrates the performance of models trained on 70% of MIMIC-II and evaluated on MIMIC-II (30% random subsample), MIMIC-III1, and MIMIC-III2. To demonstrate the effect of clinicians' knowledge, we first organized the rules into five categories according to a histogram of the absolute value of ΔRank, with $\Delta R = 0$ reflecting those rules in which clinicians agreed the most with the empirical data and 5 the least. (In this text, we use ΔRank to refer to the difference between expert-assessed risk and empirical risk and ΔR to refer to the same measure after it has been cut into five bins). The effect of building different models by serially removing rules with increasing ΔR is illustrated in Fig. 5*B*. This process can be considered as a "hard EAML," where those rules that disagree more than a certain threshold are infinitely penalized (i.e., discarded) while those below the threshold are penalized by a constant. Since these rules were selected by RuleFit using the empirical distribution on MIMIC-II, getting rid of rules adversely affects performance (area under the curve [AUC]) in the training data and in the testing set that originates from the same empirical distribution (Fig. 5*B*). A different scenario emerges when these models are tested on both MIMIC-III1 and MIMIC-III2. In this case, penalizing those rules where clinicians disagree the most with the empirical data improves performance. When only rules with $\Delta R = 0$ are left ($n = 53$ of 126 rules), however, performance decreases (Fig. 5*B*). This suggests a trade-off between using better rules to build the models (those in which clinician agree with the empirical risk) and oversimplifying the model (if only rules with $\Delta R = 0$ are used). Therefore, better results might be obtained if we acquired clinicians' answers for all 2,000 rules and not just the 126 selected by least absolute shrinkage and selection operator (LASSO). The trade-off here is time needed to collect experts' assessments.

Additionally, in Fig. 5*C* we note that models with the highest accuracy can be obtained with half the data if clinicians' answers are used to limit rules used for training: models built with rules from groups 1 and 2, i.e., where $\Delta R \leq 1$, saturate around 400 patients, while those built with all of the rules need around 800 patients. Wilcoxon tests comparing performance of models trained on 6,400 cases (saturation) using only rules with ranking difference ≤1 vs. all rules show the reduced rule set results in significantly better AUC ($W = 9$, $P = 0.00105$) and balanced accuracy ($W = 4$, $P = 0.00013$). This effect is not present if the

Table 1. The top 5% rules in which the clinician-perceived risk is greater (top) and less (bottom) than the empirical risk

	Clinician-estimated risk vs. empirical risk	ΔRank
Clinician-estimated risk > empirical risk		
1	Age = 66.15 (16.5–89.3); PaO_2/FiO_2 = 332.6 (199.0–2,304.8); HR = 84.00 (0.0–106.0); GCS = 2 (1–4); renal function = 0 (0,1)	−49
2	PaO_2/FiO_2 = 332.6 (224.0–955.0); GCS = 5 (5); age = 80.9 (74.61–101.5); renal function = 0 (0)	−48
3	GCS = 2.0 (1.0–4.0); BUN = 15.00 (2.0–24.0); age = 58.8 (16.8–75.2); PaO_2/FiO_2 = 332.6 (212.0–1,942.9); HR = 80.0 (0.00–92.0)	−47
4	HR = 80.00 (0.0–94.0); GCS = 2 (1–4); BUN = 15.0 (2.0–24.0); age = 62.7 (17.2–83.6); PaO_2/FiO_2 = 332.6 (272.0–1,942.9)	−47
5	PaO_2/FiO_2 = 332.6 (318.6–2,223.8); GCS = 5 (3–5); age = 81.2 (73.8–101.5); renal function = 0 (0)	−44
6	HR = 103.0 (93.0–171.0); GCS = 1 (1–2); BUN = 14.0 (2.0–23.0); PaO_2/FiO_2 = 345.0 (272.0–1,939.3)	−43
Clinician-estimated risk < empirical risk		
7	GCS = 5 (3–5); bilirubin = 2.7 (1.5–48.0); BUN = 35.0 (20.0–248.0)	37
8	GCS = 5 (4–5); BUN = 44.0 (27.00–272.0); BP = 91.0 (0.0–108.0)	37
9	PaO_2/FiO_2 = 496.5 (342.3–1,942.9); HR = 117.0 (107.0–171.0); BUN = 13.0 (2.0–21.0)	39
10	PaO_2/FiO_2 = 122.9 (20.0–271.4); age = 53.8 (18.3–78.4); bilirubin = 3.6 (1.6–59.7)	39
11	GCS = 5 (3–5); bilirubin = 4.0 (1.9–48.0); renal function = 1 (1–4)	55
12	Renal function = 0 (0,1); PaO_2/FiO_2 = 470.0 (336.7–2,304.8)	56

Variables likely to have driven the response are highlighted in red. Values are shown as variable = median (range).

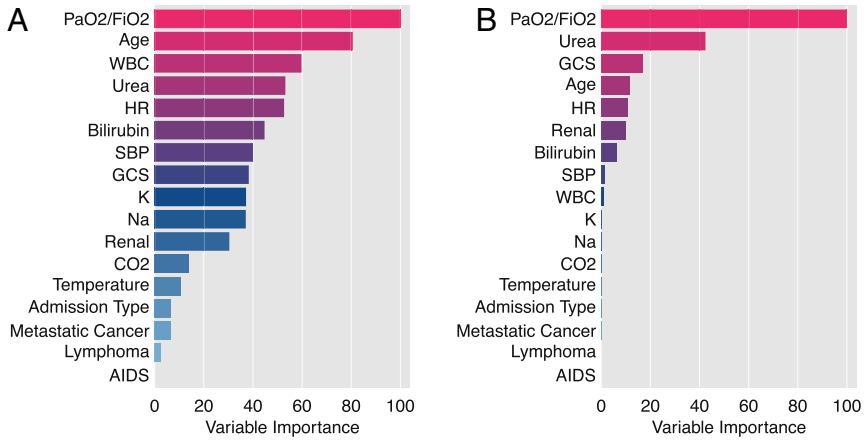

Fig. 4. Variable importance estimated using a Random Forest model predicting mortality (*A*), and clinicians' assessments (*B*). While PaO_2/FiO_2 is the most important variable in both cases, in the former case it is used to learn intubation status, while in the latter clinicians are responding based on its physiological influence on mortality.

model is trained and tested on MIMIC-III data (Fig. 5*D*). Fig. 5 *B–D* exemplifies the difficulties and limitations of selecting the best models using cross-validated errors estimated from the empirical distribution. Upon covariate shifts and data acquired at a different time (possibly reflecting new interventions and treatments, etc.), model selection using cross-validation from the empirical distribution is no longer optimal because spurious correlations found in the empirical distribution are likely to change. Since true causal knowledge does not change, our results suggest that this knowledge is being extracted from clinicians (e.g., evaluation of PaO_2/FiO_2 by clinicians). Finally, similar results can be obtained if instead of using the hard version of EAML, we use a soft version (*SI Appendix*).

Discussion

Despite increasing success and growing popularity, ML algorithms can be data inefficient and often generalize poorly to unseen cases. We have introduced EAML, the first methodology to automatically extract problem-specific clinical prior knowledge from experts and incorporate it into ML models. Related previous work had attempted to predict risk based on clinicians' assessment of individual cases using all available patient characteristics with limited success (20). Here, in contrast, we transformed the raw physiologic data into a set of simple rules and asked clinicians to assess the risk of subpopulations defined by those rules relative to the whole sample. We showed that utilizing this extracted prior knowledge allows: 1) discovery of hidden confounders and limitations of clinicians' knowledge, 2) better generalization to changes in the underlying feature distribution, 3) improved accuracy in the face of time decay, 4) training with less data, and 5) illustrating the limitations of models chosen using cross-validation estimated from the empirical distribution. We used the MIMIC dataset from the PhysioNet project (9) (10), a large dataset of intensive-care patients, to predict hospital mortality. We showed that EAML allowed the discovery of a hidden confounder (intubation) that can change the interpretation of common variables used to model ICU mortality in multiple available clinical scoring systems—APACHE (11), SAPS II (21), or SOFA (13). Google Scholar lists over 10,000 citations of PhysioNet's MIMIC dataset as of December 2018, with ∼1,600 new papers published every year. Conclusions on treatment effect or variable importance using this dataset should be taken with caution, especially since intubation status can be implicitly learned from the data, as shown in this study, even though the variable was not recorded. Moreover, we identified areas where clinicians' knowledge may need evaluation and possibly further training, such as the case where clinicians overestimated the mortality risk of old age in the absence of other strong risk factors. Further investigation is warranted to establish whether clinicians' perceived risk is negatively impacting treatment decisions.

We have built EAML to incorporate clinicians' knowledge along with its uncertainty into the final ML model. EAML is not merely a different way of regularizing a machine-learned model but is designed to extract domain knowledge not necessarily present in the training data. We have shown that incorporating this prior knowledge helps the algorithm generalize better to changes in the underlying variable distributions, which, in this case, happened after a rebuilding of the database by the PhysioNet Project. We have also demonstrated that we can train models more robust to accuracy decay with time. Preferentially using those rules where clinicians agree with the empirical data not only produces models that generalize better, but it does so with considerably less data ($n = 400$ vs. $n = 800$). This result can be of high value in multiple fields where data are scarce and/or expensive to collect. We also demonstrated the limitation of selecting models using cross-validated estimation from within the empirical distribution. We showed that there is no advantage in incorporating clinicians' knowledge if the test set is drawn from the same distribution as the training. However, when the same model was tested in a population whose variables had changed or that were acquired at a later time, including clinicians' answers improved performance and made the algorithm more data efficient.

The MIMIC dataset offered a great opportunity to demonstrate the concept and potential of EAML. A major strength of the dataset is the large number of cases, while one of the main weaknesses is that all cases originated from a single hospital. We were able to show the benefit of EAML in the context of feature coding changes and time decay (MIMIC-III1 and MIMIC-III2). However, proper application of EAML requires independent training, validation, and testing sets, ideally from different institutions. Crucially, an independent validation set is required in order to choose the best subset of rules (hard EAML) or the lambda hyperparameter (soft EAML). If the validation set has the same correlation structure between the covariates and outcome as the training set, cross-validation will choose a lambda of 0 provided there are enough data points. However, if the validation set is different from the training set, then incorporating expert knowledge will help and the tuning will result in lambda greater than 0. This is the same for any ML model training where hyperparameter tuning cannot be effectively performed

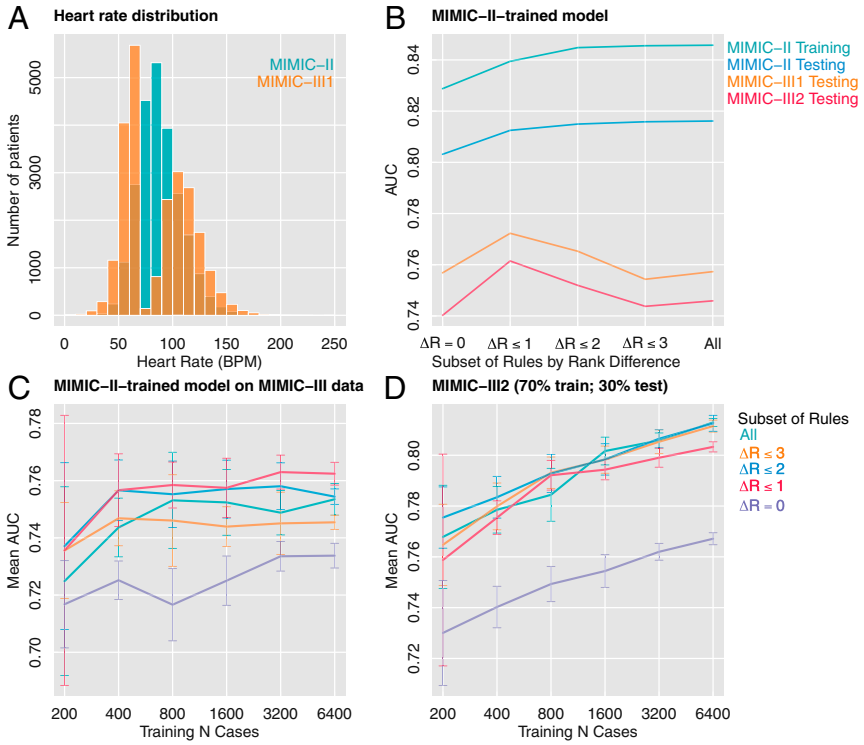

Fig. 5. Example of variable shift: heart rate distribution of the same set of patients from MIMIC-II and MIMIC-III1 (*A*). Models were trained on MIMIC-II data using different subsets of rules defined by the extent of clinicians' agreement with the empirical risk (delta ranking cut in five bins, ΔR) (*B*). Mean AUC of models trained on MIMIC-II and tested on MIMIC-III (*C*) and models trained and tested on MIMIC-III (*D*). Subsamples of different sizes were used for each subset of rules defined by ΔR to test the hypothesis that eliminating bad rules helps the algorithm train with less data. Error bars represent 1 SD across 10 stratified subsamples.

by cross-validation of the training set if that set is not representative of the whole population of interest, which is most commonly the case in clinical datasets. One of the biggest contributions of this paper is showing the risk of using a validation set that has been randomly subsampled from the empirical distribution and as such contains the same correlations as the training data. Our team is preparing a multiinstitutional EAML study to optimize the algorithm for real-world applications.

Finally, this work also has implications on the interpretability and quality assessment of ML algorithms. It is often considered that a trade-off exists between interpretability and accuracy of ML models (22, 23). However, as shown by Friedman and Popescu (24), rule ensembles, and therefore EAML, are on average more accurate than Random Forest and slightly more accurate than Gradient Boosting in a variety of complex problems. EAML builds on RuleFit to address the accuracy–interpretability trade-off in ML and allows one to examine all of the model's rule ahead of deployment, which is essential to building trust in predictive models.

Methods

A more complete description of the study methods is available in *SI Appendix*. Briefly, we used the publicly available MIMIC ICU dataset from the PhysioNet project to predict in-hospital mortality. The MIMIC dataset includes two releases: MIMIC-II, collected at BIDMC between 2001 and 2008 (9), and MIMIC-III (10), which includes the MIMIC-II cases after recoding of some variables (which resulted in distribution shifts) plus new cases treated between 2008 and 2012. We split the data in four groups: 1) MIMIC-II training (70% of MIMIC-II stratified on outcome); 2) MIMIC-II testing (remaining 30% of MIMIC-II); 3) MIMIC-III1 (MIMIC-II cases after recoding); and 4) MIMIC-III2 (new cases collected after 2008 not present in MIMIC-II). The 17 input features consisted of demographics and clinical and physiological variables included in common ICU risk scoring systems.

The RuleFit procedure (24) was used to derive 126 decision rules made up of three to five input variables that predict mortality. These rules represent a transformation of the input variables to a Boolean matrix (i.e., True/False). For example, the rule "Age > 75 & systolic blood pressure < 80 & Glasgow Coma Scale < 10" will have a value of "1" for all patients that match each of these conditions and "0" otherwise, thus defining a subpopulation within the full sample. The RuleFit-derived rules were uploaded to a web application (http://www.mediforest.com/). Fifteen hospitalists and ICU clinicians were asked to assess the relative mortality risk of patients belonging to the subgroup defined by each rule relative to the whole population by selecting one of five possible responses: highly decrease, 1; moderately decrease, 2; no effect, 3; moderately increase, 4; and highly increase, 5. Rules were ranked based on the empirical risk of their respective subpopulations (Rank_e) and by the mean clinician-assessed risk (Rank_p). The difference $\Delta\text{Rank} = \text{Rank}_p - \text{Rank}_e$ was calculated to represent the extent of agreement between the empirical data and the expert assessments and was used 1) to identify problems in the training data and 2) to regularize the final EAML model by penalizing rules with higher disagreement. All analysis and visualization were performed using the rtemis machine learning library (25).

Data Availability. The software used in this study is available on GitHub at https://github.com/egenn/rtemis. The code used to perform this study along with the rankings obtained from clinicians is available at https://github.com/egenn/EAML_MIMIC_ICUmortality. The MIMIC dataset (9) can be obtained after registration with the PhysioNet project (https://physionet.org/).

ACKNOWLEDGMENTS. We thank two anonymous reviewers for their constructive feedback. Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering under Award K08EB026500 (G.V.), by the National Institute of Allergy and Infectious Diseases under Award 5R01AI074345-09 (M.J.v.d.L.), by the National Center for Advancing Translational Sciences through University of California, San Francisco-Clinical & Translational Science Institute Grant UI1TR001872 (G.V.), and by the Wicklow AI in Medicine Research Initiative at the University of San Francisco Data Institute (L.G.R. and Y.I.).

1. D. B. Lenat, M. Prakash, M. Shepherd, CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine* **6**, 65 (1985).
2. E. W. Steyerberg *et al*.; PROGRESS Group, Prognosis Research Strategy (PROGRESS) 3: Prognostic model research. *PLoS Med.* **10**, e1001381 (2013).
3. A. D. Hingorani *et al*.; PROGRESS Group, Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ* **346**, e5793 (2013).
4. G. F. Cooper *et al*., Predicting dire outcomes of patients with community acquired pneumonia. *J. Biomed. Inform.* **38**, 347–366 (2005).
5. S. Mullainathan, Z. Obermeyer, Does machine learning automate moral hazard and error? *Am. Econ. Rev.* **107**, 476–480 (2017).
6. P. Rajpurkar *et al*., CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv:1711.05225v3 (25 December 2017).
7. J. R. Zech *et al*., Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
8. J. R. Zech *et al*., Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
9. M. Saeed *et al*., Multiparameter intelligent monitoring in intensive care II: A public-access intensive care unit database. *Crit. Care Med.* **39**, 952–960 (2011).
10. A. E. W. Johnson *et al*., MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035–160039 (2016).
11. W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, D. E. Lawrence, APACHE-acute physiology and chronic health evaluation: A physiologically based classification system. *Crit. Care Med.* **9**, 591–597 (1981).
12. J. R. Le Gall *et al*., A simplified acute physiology score for ICU patients. *Crit. Care Med.* **12**, 975–977 (1984).
13. J. L. Vincent *et al*., The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Med.* **22**, 707–710 (1996).
14. R. Pirracchio *et al*., Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. *Lancet Respir. Med.* **3**, 42–52 (2015).
15. J. I. F. Salluh, M. Soares, ICU severity of illness scores: APACHE, SAPS and MPM. *Curr. Opin. Crit. Care* **20**, 557–565 (2014).
16. A. E. Johnson, D. J. Stone, L. A. Celi, T. J. Pollard, The MIMIC code repository: Enabling reproducibility in critical care research. *J. Am. Med. Inform. Assoc.* **25**, 32–39 (2018).
17. K. Beier *et al*., Elevation of blood urea nitrogen is predictive of long-term mortality in critically ill patients independent of "normal" creatinine. *Crit. Care Med.* **39**, 305–313 (2011).
18. D. K. Rajan, Z. J. Haskal, T. W. I. Clark, Serum bilirubin and early mortality after transjugular intrahepatic portosystemic shunts: Results of a multivariate analysis. *J. Vasc. Interv. Radiol.* **13**, 155–161 (2002).
19. J. M. Engel *et al*., Outcome prediction in a surgical ICU using automatically calculated SAPS II scores. *Anaesth. Intensive Care* **31**, 548–554 (2003).
20. N. White, F. Reid, A. Harris, P. Harries, P. Stone, A systematic review of predictions of survival in palliative care: How accurate are clinicians and who are the experts? *PLoS One* **11**, e0161407 (2016).
21. J. R. Le Gall, S. Lemeshow, F. Saulnier, A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *JAMA* **270**, 2957–2963 (1993).
22. G. Valdes *et al*., MediBoost: A patient stratification tool for interpretable decision making in the era of precision medicine. *Sci. Rep.* **6**, 37854 (2016).
23. R. Caruana *et al*., "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission" in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining–KDD '15* (ACM Press, 2015), pp. 1721–1730.
24. J. H. Friedman, B. E. Popescu, Predictive learning via rule ensembles. *Ann. Appl. Stat.* **2**, 916–954 (2008).
25. E. D. Gennatas, Towards precision psychiatry: Gray matter development and cognition in adolescence. Publicly accessible Penn dissertations 2302. https://repository.upenn.edu/edissertations/2302. Accessed 29 April 2019.